IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 13, 2021, accepted February 6, 2022, date of publication February 28, 2022, date of current version March 8, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3155233

Graph Embedding With Data Uncertainty

FIRAS LAAKOM[1], JENNI RAITOHARJU[2], (Member, IEEE), NIKOLAOS PASSALIS[3], ALEXANDROS IOSIFIDIS[4], (Senior Member, IEEE), AND MONCEF GABBOUJ[1], (Fellow, IEEE)

[1]Faculty of Information Technology and Communication Sciences, Tampere University, 33100 Tampere, Finland
[2]Programme for Environmental Information, Finnish Environment Institute, 40500 Jyväskylä, Finland
[3]Department of Informatics, Aristotle University of Thessaloniki, 541 24 Thessaloniki, Greece
[4]Department of Electrical and Computer Engineering, Aarhus University, 8000 Aarhus, Denmark

Corresponding author: Firas Laakom (firas.laakom@tuni.fi)

This work was supported by NSF-Business Finland Center for Visual and Decision Informatics (CVDI) Project Advanced Machine Learning for Industrial Applications (AMALIA). The work of Jenni Raitoharju was supported by the Academy of Finland under Project 324475.

ABSTRACT Spectral-based subspace learning is a common data preprocessing step in many machine learning pipelines. The main aim is to learn a meaningful low dimensional embedding of the data. However, most subspace learning methods do not take into consideration possible measurement inaccuracies or artifacts that can lead to data with high uncertainty. Thus, learning directly from raw data can be misleading and can negatively impact the accuracy. In this paper, we propose to model artifacts in training data using probability distributions; each data point is represented by a Gaussian distribution centered at the original data point and having a variance modeling its uncertainty. We reformulate the Graph Embedding framework to make it suitable for learning from distributions and we study as special cases the Linear Discriminant Analysis and the Marginal Fisher Analysis techniques. Furthermore, we propose two schemes for modeling data uncertainty based on pair-wise distances in an unsupervised and a supervised contexts.

INDEX TERMS Graph embedding, subspace learning, dimensionality reduction, uncertainty estimation, spectral learning.

I. INTRODUCTION

With the advancement of data collection processes, high dimensional data are available for applying machine learning approaches. However, the impracticability of working in high dimensional spaces due to the *curse of dimensionality* and the realization that the data in many problems reside on manifolds with much lower dimensions than those of the original space, has led to the development of spectral-based subspace learning (SL) techniques. Spectral-based methods rely on the eigenanalysis of Scatter matrices. SL aims at determining a mapping of the original high-dimensional space into a lower-dimensional space preserving properties of interest in the input data. This mapping can be obtained using unsupervised methods, such as Principal Component Analysis (PCA) [1], [2], or supervised ones, such as Linear Discriminant Analysis (LDA) [3] and Marginal Fisher Analysis (MFA) [4]. Despite the different motivations of these spectral-based methods, a general formulation known as Graph Embedding was introduced in [4] to unify them within a common framework.

For low-dimensional data, where dimensionality reduction is not needed and classification algorithms can be applied directly, many extensions modeling input data inaccuracies have recently been proposed [5], [6]. In [6], data points are replaced by probability distributions modeling the artifacts and an SVM classifier was extended to operate on data distributions. However, for high dimensional data, dimensionality reduction is needed. If the provided data is exposed to measurement inaccuracies or artifacts, learning directly from it can lead to a biased or erroneous embedding of the high dimensional data [5], [6]. Traditional methods, such as LDA and MFA do not take this into consideration. Extensions of some SL methods taking into account the presence of outliers and noise in the data were proposed to tackle for this problem, such as the methods in [9], [10] for LDA, and the method in [11] for PCA.

In this paper, we propose a novel spectral-based subspace learning framework, called Graph Embedding with Data Uncertainty (GEU), in which input data uncertainties are

The associate editor coordinating the review of this manuscript and approving it for publication was Zhiwei Gao.

taken into consideration. Instead of relying on the training data directly, we model each data point by a multivariate Gaussian distribution centered at the position of the original measurement and having a covariance matrix accounting for its uncertainty. To this end, we reformulate the Graph Embedding framework to operate on distributions at individual data point level allowing us to determine a mapping from the input data space into a lower-dimensional space via optimizing some properties of interest defined over these distributions. The outcome is a more robust data embedding scheme. As special cases of the proposed framework formulations, we investigate extensions of LDA and MFA techniques within the proposed GEU framework. We refer to these as GEU-LDA and GEU-MFA, respectively. An example of the decision boundaries obtained by using the original MFA, MFA with augmented data, and GEU-MFA on 2-D synthetic data forming two classes is illustrated in Figure 1. The incorporation of data uncertainty shifts the decision boundary of the original approach. We note that by using more augmented data the decision boundary of MFA shifts toward the GEU-MFA.

Furthermore, we theoretically show that under the proposed GEU framework, the rank of matrices involved in the optimization problem, i.e., the scatter matrices, increases compared to the original methods. As a result, methods formulated under the proposed framework lead to an increased number of projection directions. This is because the covariances employed to model the uncertainty at the level of the individual data point introduce a regularization term to both scatter matrices. Thus, an indirect advantage of formulating traditional SL methods, such as LDA, under the proposed framework is that it allows for addressing the small sample size problem [12], even for problems formed by two classes.

Although the focus in this paper in on LDA and MFA, the proposed GEU framework operating on generic graph structures can directly be used to obtain robust solutions for other SL methods formulated under the Graph Embedding framework. The contributions of the paper are as follows:

- We propose a novel spectral-based subspace learning framework which takes into consideration uncertainties in the input data.
- We reformulate the Graph Embedding framework to operate on distributions at individual data points. In this way, we provide a generic approach for accounting for data uncertainties in a multitude of SL methods expressed under the Graph Embedding framework.
- We study as special cases of the proposed framework GEU-LDA and GEU-MFA, and we theoretically show that considering uncertainty leads to an increased number of projection directions.
- We propose two schemes to model uncertainty of each sample based on pair-wise distances of data points in the original space.

The remainder of the paper is organized as follows. Section II provides a brief review of the related work. Section III describes in detail the proposed GEU framework. Section IV provides the conducted experimental analysis, and Section V concludes our work.

II. RELATED WORK

A. GRAPH EMBEDDING

Graph Embedding [4], [13], [14] is a general framework encapsulating several SL methods as special cases. Data points are modeled as vertices of two graph structures, namely an intrinsic graph expressing data relationships to be emphasized and a penalty graph expressing data relationships to be suppressed. Using such intrinsic and penalty graphs, the optimization problems of SL methods, such as LDA, PCA, and MFA, can be formulated.

Given a set of data points followed by the corresponding class labels $\{(\mathbf{x}_i, c_i)\}_{i=1}^{N}$, where $\mathbf{x}_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$, the goal in Graph Embedding is to determine a mapping which maps $\mathbf{x}_i$ to a lower dimensional representation $\mathbf{y}_i \in \mathbb{R}^d$, $d < D$. This is achieved by forming a weighted (intrinsic) graph $G = \{\mathbf{X}, \mathbf{W}\}$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ is the vertex set and $\mathbf{W} \in \mathbb{R}^{N \times N}$ the graph weight matrix whose elements encode the pair-wise relationships between the graph vertices $\mathbf{x}_i$. Furthermore, a penalty graph $G^p = \{\mathbf{X}, \mathbf{W}^p\}$ can be defined on the same graph vertices, whose weight matrix $\mathbf{W}^p \in \mathbb{R}^{N \times N}$ expresses pair-wise relationships to be penalized. For example, the goal can be to emphasize connections of points within the same class, i.e., to have them close to each other in the embedding space, and suppress the connections of points from different classes, i.e., to have them distant in the embedding space.

The graph preserving criterion is formulated as follows:

$$\mathbf{y}^* = \underset{\mathbf{y}^T\mathbf{B}\mathbf{y}=m}{\arg\min} \sum_{i \neq j} (y_i - y_j)^2 \mathbf{W}_{ij}, \tag{1}$$

where $\mathbf{y} = [y_1, \ldots, y_N]^T$, $y_i \in \mathbb{R}$ is a 1-D mapping of $\mathbf{x}_i$, m is a constant and $\mathbf{B}$ can be defined as a constraint matrix, e.g., $\mathbf{B} = \mathbf{I}$ to enforce orthogonality constraints, or as a scatter matrix based on the Laplacian of the penalty graph. The value of m depends on the approach used. For example, it is set to 1 for LDA and MFA. For a linear data mapping, i.e., $\mathbf{y} = \mathbf{X}^T\mathbf{v}$, where $\mathbf{v} \in \mathbb{R}^D$ is a unitary projection vector mapping $\mathbf{x}_i \in \mathbb{R}^D$ to $y_i \in \mathbb{R}$, Eq. (1) can be rewritten as follows:

$$\mathbf{v}^* = \underset{\mathbf{v}^T\mathbf{X}\mathbf{B}\mathbf{X}^T\mathbf{v}=m}{\arg\min} \mathbf{v}^T\mathbf{X}\mathbf{L}\mathbf{X}^T\mathbf{v}, \tag{2}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix with $\mathbf{D}$ being the diagonal degree matrix having elements $\mathbf{D}_{ii} = \sum_{j \neq i} \mathbf{W}_{ij}$, and $\mathbf{B} = \mathbf{X}\mathbf{L}^p\mathbf{X}^T = \mathbf{X}(\mathbf{D}^p - \mathbf{W}^p)\mathbf{X}^T$. In this case, the solution of the optimization problem in Eq. (2) is given by solving the generalized eigenvalue decomposition problem

$$\left(\mathbf{X}\mathbf{L}\mathbf{X}^T\right)\mathbf{v} = \lambda\left(\mathbf{X}\mathbf{L}^p\mathbf{X}^T\right)\mathbf{v} \tag{3}$$

and keeping the eigenvector corresponding to the smallest (positive) eigenvalue. To obtain more than one projection direction, the corresponding projection matrix $\mathbf{V} \in \mathbb{R}^{D \times d}$ is formed by the eigenvectors corresponding to the d smallest eigenvalues.

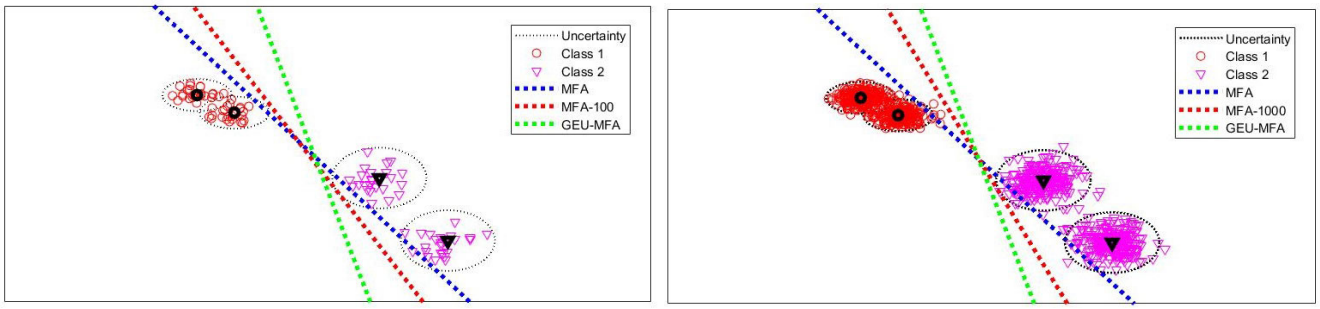

FIGURE 1. The decision functions obtained by using MFA, GEU-MFA and MFA applied on augmented data by 100 samples, i.e., MFA-100 (left) and 1000 samples, i.e. MFA-1000 (right).

Specific selections of $\mathbf{W}$ and $\mathbf{W}^p$ lead to different subspace learning methods. For LDA, the within-class scatter and the between-class scatter matrices are given by

$$\mathbf{S}_w = \mathbf{X}\left(\mathbf{I} - \sum_{c=1}^{C} \frac{1}{N_c}\mathbf{e}^c\mathbf{e}^{cT}\right)\mathbf{X}^T, \tag{4}$$

$$\mathbf{S}_b = \mathbf{X}\left(\sum_{c=1}^{C} N_c\left(\frac{1}{N_c}\mathbf{e}^c - \frac{1}{N}\mathbf{e}\right)\left(\frac{1}{N_c}\mathbf{e}^c - \frac{1}{N}\mathbf{e}\right)^T\right)\mathbf{X}^T, \tag{5}$$

where C is the number of classes, N_c is the cardinality of class c, $\mathbf{e} \in R^N$ is the vector with all elements equal to 1, and $\mathbf{e}^c \in R^N$ is a vector with the elements corresponding to data points of class c equal to one and the rest equal to zero. Thus, LDA can be formulated in the Graph Embedding framework by using the graph weight matrices

$$\mathbf{W}_{ij} = \begin{cases} \frac{1}{N_{c_i}}, & \text{if } c_i = c_j \text{ and } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$\mathbf{W}^p_{ij} = \begin{cases} \frac{1}{N} - \frac{1}{N_{c_i}}, & \text{if } c_i = c_j \text{ and } i \neq j \\ \frac{1}{N}, & \text{otherwise} \end{cases} \tag{7}$$

where N_{c_i} is the cardinality of the class, which $\mathbf{x}_i$ belongs to. MFA is formulated by using the graph weight matrices

$$\mathbf{W}_{ij} = \begin{cases} 1, & \text{if } i \in N^+_{k1}(j) \text{ or } j \in N^+_{k1}(i) \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

$$\mathbf{W}^p_{ij} = \begin{cases} 1, & \text{if } (i,j) \in P_{k_2}(c_i) \text{ or } (i,j) \in P_{k_2}(c_j) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $N^+_{k1}(j)$ is the set of the k_1 nearest neighbors of the $\mathbf{x}_j$ in the same class, and $P_{k_2}(c)$ is the set the k_2 nearest pairs among the set $\{(i,j), \mathbf{x}_i \in c, \mathbf{x}_j \notin c\}$. Here, we should note that several other methods which employ pair-wise similarity/distance measures, e.g. [8], [13], [15]–[20], can be formulated using the Graph Embedding framework.

B. LEARNING WITH UNCERTAINTY

Research in uncertainty has gained a lot of attention lately in many branches of science [21], [22], since data can be subject to measurement inaccuracies and artifacts. Taking this into consideration in the data modeling and learning process is critical for building robust models. Exploiting uncertainty in machine learning has been studied from many different viewpoints. Methods dealing with uncertainty can be grouped into two different categories: sample-wise uncertainty modeling and feature-wise uncertainty modeling.

In sample-wise uncertainty, the noise is modeled at the sample level. The main assumption in such methods is that few training data points are outliers and thus they need to be suppressed or partially suppressed to not affect the solution of the subsequent processing steps. Various robust extensions of SL methods have been proposed to reduce the sensitivity of a classifier to outliers [7], [9]–[11], [23]–[26]. In [23] and [24] for example, robust extensions of LDA were proposed by reducing the sensitivity of the model to outliers.

In feature-wise uncertainty, the noise is modeled at the data dimension level. The main assumption in such methods is that certain data dimensions are corrupted by noise. This type of noise modeling was employed to extend SVM in [6]. For SL, feature-wise uncertainty is used in [9], where a robust extension of LDA is proposed. Instead of using point estimates of speech data, a probabilistic description based on Gaussian distributions at the individual data point level are used as inputs to LDA. In our work, we use a similar uncertainty modeling. However, we note two key differences: (i) The approach in [9] is restricted to the LDA method, whereas our work is based on the Graph Embedding framework formulation of SL and, thus, it is more general and can be used in several approaches, e.g., LDA, PCA, and MFA. ii) The uncertainty estimation approach in [9] is restricted to speech data and in [6] is restricted to image data, whereas we propose two schemes to model the uncertainty of each sample based on pair-wise distances of data points in the original space. Thus, our approach of modeling uncertainty can be used for any type of data.

III. GRAPH EMBEDDING WITH DATA UNCERTAINTY

Let us denote by $\{\underline{y}_i\}_{i=1}^N$ a set of the random Gaussian variables expressing the low-dimensional representations of the input data $\mathbf{x}_i$, $i = 1, \ldots, N$. We express the graph preserving

criterion using $\underline{y}_i$ as follows:

$$\underline{\mathbf{y}}^* = \underset{\mathbb{E}(\underline{\mathbf{y}}^T\mathbf{B}\underline{\mathbf{y}})=m}{\arg\min} \sum_{i\neq j} \mathbb{E}\left((\underline{y}_i - \underline{y}_j)^2\right)\mathbf{W}_{ij}, \tag{10}$$

where $\mathbb{E}(\cdot)$ denotes the expectation operator. For a Gaussian uncertainty, i.e., $\underline{y}_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, the pair-wise distances $\underline{z}_{ij}$ between $\underline{y}_i$ and $\underline{y}_j$ are also random variables following a Gaussian distribution

$$\underline{z}_{ij} = \underline{y}_i - \underline{y}_j \sim \mathcal{N}(\mu_i - \mu_j, \sigma_i^2 + \sigma_j^2). \tag{11}$$

Thus, the expectation term in Eq. (10) can then be rewritten as follows:

$$\begin{aligned}\mathbb{E}((\underline{y}_i - \underline{y}_j)^2) &= \mathbb{E}(\underline{z}_{ij}^2) = \mathbb{E}(\underline{z}_{ij})^2 + Var(\underline{z}_{ij}) \\ &= (\mu_i - \mu_j)^2 + (\sigma_i^2 + \sigma_j^2).\end{aligned} \tag{12}$$

By substituting Eq. (12) to Eq. (10), we get

$$\begin{aligned}\underline{\mathbf{y}}^* &= \underset{\mathbb{E}(\underline{\mathbf{y}}^T\mathbf{B}\underline{\mathbf{y}})=m}{\arg\min} \sum_{i\neq j} \mathbb{E}\left((\underline{y}_i - \underline{y}_j)^2\right)\mathbf{W}_{ij} \\ &= \underset{\mathbb{E}(\underline{\mathbf{y}}^T\mathbf{B}\underline{\mathbf{y}})=m}{\arg\min} \sum_{i\neq j} \left((\mu_i - \mu_j)^2 + (\sigma_i^2 + \sigma_j^2)\right)\mathbf{W}_{ij}\end{aligned} \tag{13}$$

The first term of the summation is equivalent to the original Graph Embedding and depends on $\mathbb{E}(\mathbf{y})$, i.e., the expectation of $\mathbf{y}$:

$$\sum_{i\neq j}(\mu_i - \mu_j)^2\mathbf{W}_{ij} = 2\,\mathbb{E}(\mathbf{y})^T\mathbf{L}\,\mathbb{E}(\mathbf{y}). \tag{14}$$

By defining

$$\boldsymbol{\sigma} = \left[\sqrt{\sigma_1^2}, \ldots, \sqrt{\sigma_i^2}, \sqrt{\sigma_n^2}\right],$$

the second term in the summation can be expressed as follows:

$$\sum_{i\neq j}(\sigma_i^2 + \sigma_j^2)\mathbf{W}_{ij} = 2\boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma}. \tag{15}$$

Thus, using Eq. (14) and Eq. (15), our new graph preserving criterion is given as follows:

$$\underline{\mathbf{y}}^* = \underset{\mathbb{E}(\underline{\mathbf{y}}^T\mathbf{B}\underline{\mathbf{y}})=m}{\arg\min}\ \mathbb{E}(\mathbf{y})^T\mathbf{L}\,\mathbb{E}(\mathbf{y}) + \boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma}. \tag{16}$$

For a linear data mapping $\underline{\mathbf{y}} = \underline{\mathbf{X}}^T\mathbf{v}$ and modeling each data point in the input space using a Gaussian distribution, i.e., $\underline{\mathbf{x}}_i \sim \mathcal{N}(\boldsymbol{\mu}_i^x, \boldsymbol{\Sigma}_i^x)$, $\underline{y}_i = \mathbf{v}^T\underline{\mathbf{x}}_i$ corresponds to a linear projection of a Gaussian, which is a Gaussian distribution $y_i \sim \mathcal{N}(\mu_i^y, (\sigma_i^y)^2)$ with $\mu_i^y = \mathbf{v}^T\boldsymbol{\mu}_i^x$ and $(\sigma_i^y)^2 = \mathbf{v}^T\boldsymbol{\Sigma}_i^x\mathbf{v}$. Thus, the second term in Eq. (16) can be written as follows:

$$\boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma} = \mathbf{v}^T\left(\sum_i \mathbf{D}_{ii}\boldsymbol{\Sigma}_i^x\right)\mathbf{v}. \tag{17}$$

The equality in Eq. (17) follows from: $\boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma} = \sum_i \sigma_i \sum_j(\mathbf{D}_{ij}\sigma_j)$. Since $\mathbf{D}$ is diagonal, $\sum_j(\mathbf{D}_{ij}\sigma_j) = \mathbf{D}_{ii}\sigma_i$. Thus, $\boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma} = \sum_i \sigma_i^2\mathbf{D}_{ii}$. In addition, $\sigma_i^2 = \mathbf{v}^T\boldsymbol{\Sigma}_i^x\mathbf{v}$, thus $\boldsymbol{\sigma}^T\mathbf{D}\boldsymbol{\sigma} = \mathbf{v}^T\left(\sum_i \mathbf{D}_{ii}\boldsymbol{\Sigma}_i^x\right)\mathbf{v}$.

Based on the above, the final form of Eq. (16) is

$$\mathbf{v}^* = \underset{\mathbb{E}(\mathbf{v}^T\underline{\mathbf{X}}\mathbf{B}\underline{\mathbf{X}}^T\mathbf{v})=m}{\arg\min}\ \mathbf{v}^T\left(\mathbb{E}(\mathbf{X})^T\mathbf{L}\,\mathbb{E}(\mathbf{X}) + \sum_i \mathbf{D}_{ii}\boldsymbol{\Sigma}_i^x\right)\mathbf{v}. \tag{18}$$

Following a derivation similar to the above, we note that a similar graph preserving criterion can be formulated with the constraint:

$$\mathbf{B} = \left(\mathbb{E}(\mathbf{X})^T\mathbf{L}^p\,\mathbb{E}(\mathbf{X}) + \sum_i \mathbf{D}_{ii}^p\boldsymbol{\Sigma}_i^x\right). \tag{19}$$

The solution of the optimization problem in Eq. (18) is given by solving the following eigenvalue decomposition problem

$$\left(\mathbb{E}(\mathbf{X})^T\mathbf{L}\,\mathbb{E}(\mathbf{X}) + \sum_i \mathbf{D}_{ii}\boldsymbol{\Sigma}_i^x\right)\mathbf{v} = \lambda\mathbf{B}\mathbf{v} \tag{20}$$

and keeping the eigenvector corresponding to the smallest (positive) eigenvalue. To obtain more than one projection directions, the corresponding projection matrix $\mathbf{V} \in \mathbb{R}^{D\times d}$ is formed by the eigenvectors corresponding to the d smallest eigenvalues.

From Eq. (18), we can observe that when uncertainty is not used, i.e., by having $\boldsymbol{\Sigma}_i^x$ equal to zero, the Gaussian distributions $\underline{\mathbf{x}}_i$ become equivalent to Dirac function. Hence, in that case, Eq. (18) becomes equivalent to Eq. (2) and the solution of the proposed approach is equivalent to that of the original Graph Embedding framework. It should be noted that, as explained above, the projected data $\underline{y_i}^*$ obtained for each data point $\mathbf{x}_i$ is also a random variable characterised by the mean $\mathbb{E}(y_i) = \mathbf{v}^T\boldsymbol{\mu}_i^x$ and variance $\sigma_i^y = \mathbf{v}^T\boldsymbol{\mu}_i^x\mathbf{v}$. One can use this additional information for the projected data or only employ the first order approximation, i.e., the mean $\mathbb{E}(y_i)$, as the final projection of the original sample $\mathbf{x}_i$. In this paper, we use the latter in the classification step.

A. EXPLOITING DATA UNCERTAINTY AS A FORM OF REGULARIZATION

By observing the eigenanalysis problem in Eq. (3), we can see that the number of projection directions which can be defined by the Graph Embedding framework depends on the underlying structure of the intrinsic and penalty graphs. That is, the maximal number of projection directions is upper bounded by the smallest rank of matrices $\mathbf{X}\mathbf{L}\mathbf{X}^T$ and $\mathbf{X}\mathbf{L}^p\mathbf{X}^T$. For example, when expressing LDA through Graph Embedding , the maximal number of projection directions is equal to the rank of $\mathbf{S}_b = \mathbf{X}\mathbf{L}^p\mathbf{X}^T$, i.e., $\min(D, C-1)$, where C is the number of classes. This restricts the number of meaningful projection directions that can be defined, leading to the extreme case of only one projection direction for binary problems. In order to solve the generalized eigenanalysis problem in Eg. (3), a regularized version $\tilde{\mathbf{S}}_b = \mathbf{X}\mathbf{L}^p\mathbf{X}^T + \epsilon\mathbf{I}$ with $\epsilon > 0$ is used, because the original $\mathbf{S}_b$ is singular. However, this regularization procedure simply shifts the eigen-spectrum of $\mathbf{S}_b$ from

λ_i to $\tilde{\lambda}_i = \lambda_i + \epsilon \geq 0,\ i = 1, \ldots, D$) and has no data-driven intuition.

From Eq. (20) we can see that both matrices involved in the generalized eigenanalysis problem of the proposed approach are strictly positive definite. That is, the additional terms $\sum_i \mathbf{D}_{ii}\boldsymbol{\Sigma}_i^x$ and $\sum_i \mathbf{D}_{ii}^p\boldsymbol{\Sigma}_i^x$ introduced to the scatter matrices defined over the intrinsic and penalty graphs act as regularization terms leading to full-rank matrices. This is due to that the Gaussian distribution covariance matrix, $\boldsymbol{\Sigma}_i^x$, is a strictly positive-definite matrix. Hence, the introduction of the proposed approach to model uncertainty at the individual data point level results in an intuitive regularization procedure, increasing the number of projection directions. This allows avoiding the small sample size problem of LDA [12], i.e., in standard LDA the number of projection direction is theoretically limited by the number of classes. Our approach solves this problem and provides more projection directions.

B. UNCERTAINTY ESTIMATION

In the proposed GEU framework, we encode the uncertainty of each individual data point by a Gaussian distribution centered at the position of the data point and having a variance which needs to be appropriately determined to reflect the properties of the problem at hand. However, data is commonly available without such uncertainty information. We propose two schemes for defining such a variance estimate based on pair-wise distance between data points in the unsupervised and the supervised settings.

Each sample $\underline{\mathbf{x}}_i$ is defined by its mean $\mathbb{E}(\underline{\mathbf{x}}_i) = \mathbf{x}_i$ for both techniques and its covariance $\boldsymbol{\Sigma}_i$ defined as follows:

$$\boldsymbol{\Sigma}_i = \sigma\ \mathrm{diag}\Big(\mathbf{x}_i - \mathbf{x}_{i^*}\Big)^2, \tag{21}$$

where σ is a constant, diag$(\cdot)$ is the diagonal operator, and $\mathbf{x}_{i^*}$ is the closest data point to $\mathbf{x}_i$ in the admissible set. For the unsupervised case, the admissible set is composed of all the training data except $\mathbf{x}_i$ and for the supervised case the admissible set is composed of all the training data except $\mathbf{x}_i$ and having the same class as $\mathbf{x}_i$. This makes our approach generic and suitable for all learning scenarios, i.e., supervised training in the presence of label information or unsupervised in the absence.

In Figure 2, we illustrate how $\mathbf{x}_{i^*}$ in Eq. 21 is selected in both variants. We consider a binary case example. For first class, we have three samples, x_1, x_2, and x_3. For the second class, we have two samples x_4 and x_5. For example, for x_1, the closest point to it is x_5 but the closest point within the same class is x_2. Thus, for x_1, x_{i^*} in the unsupervised case is x_5 and in the supervised case is x_2. For x_3, x_{i^*} in both the unsupervised and supervised cases is x_2 as that it is the closest point to it and it belongs to same class 1. Similarly, we can deduce the neighbor for each of the remaining points in both cases.

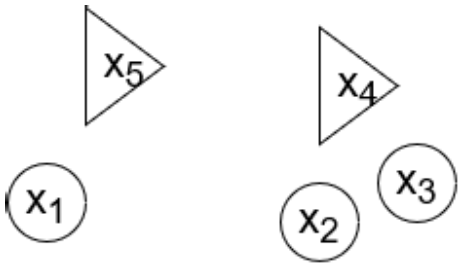

FIGURE 2. 2D illustrative example on how the closest samples considered in our uncertainty estimation can be different in the supervised and unsupervised variants. Samples x_1, x_2, and x_3 belong to class 1, x_4 and x_5 belong to the second class. The closest sample to x_1 is x_5 in the unsupervised case, but x_2 in the supervised case.

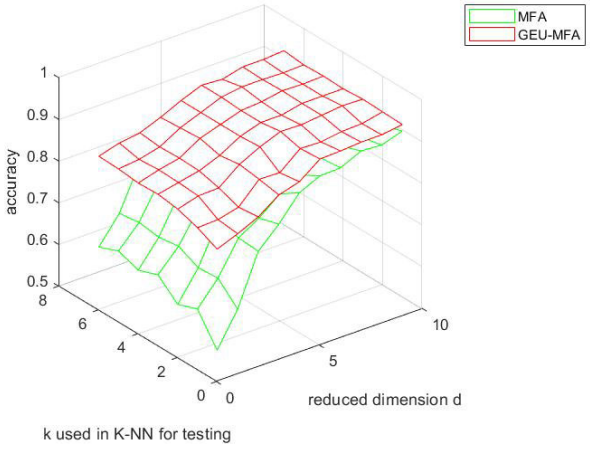

FIGURE 3. Performance evaluation of MFA and GEU-MFA on Breast Cancer Wisconsin dataset for different combination of d, the dimension of reduced space, and k used in k-NN.

IV. EXPERIMENTS AND ANALYSIS

In this section, we study as special cases of the proposed framework the traditional subspace learning techniques LDA and MFA using our learning paradigm. For all testing scenarios, we rely on Nearest Neighbor for the classification. For the evaluation, we use three different datasets:

- Breast Cancer Wisconsin dataset [27]: It is a binary classification dataset composed of 569 samples with 32 features. An explicit uncertainty estimate is proposed in [6]. We use a random 5-fold split for the evaluation of different approaches. We keep the folds fixed for the different methods.
- Cifar2: We use two classes, ''cat'' and ''dog'', from the original Cifar10 [28]. We randomly sample 900 images per class for the training. For the testing, we use the original test set of Cifar10 for both classes. To reduce the computational complexity, we first apply Bag of Visual Words (BoVW) using the SIFT descriptors to get a 400-dimensional representations of the original data.
- Extended Yale B Face Database [29]: It contains 38 subjects and each subject provides 64 face images with

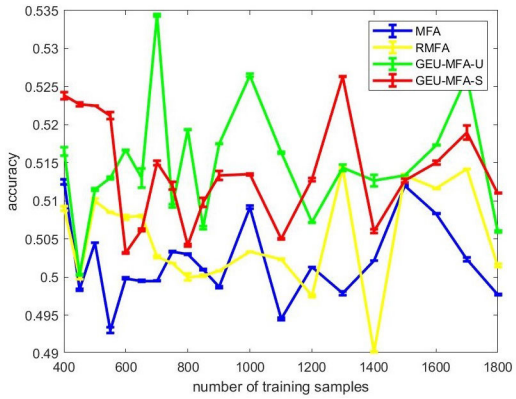

FIGURE 4. Average accuracy and variances of MFA, RMFA, GEU-MFA-U, and GEU-MFA-S on Cifar2 for different training set sizes.

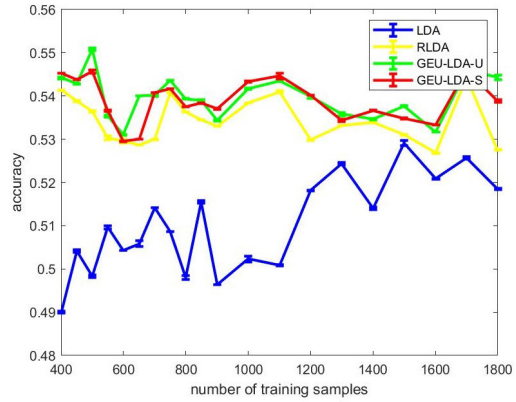

FIGURE 5. Average accuracy and variances of LDA, RLDA, GEU-LDA-U, and GEU-LDA-S on Cifar2 for different training set sizes.

different illumination conditions. Similar to [23], we crop each image and convert it to a 32 by 32 gray image. Then, PCA is used to extract a 148 feature vector per sample.

For all experiments, we cross-validate for the value of σ from $\{0.001, 0.1, 0.2, 0.4, 0.8, 1, 2\}$ and for the projection space dimension d from $\{1, 2, 4, 8\}$. We denote the supervised and unsupervised variants of uncertainty estimation with S and U, respectively.

A. MFA

MFA is a SL technique which characterizes the intraclass compactness in the intrinsic graph and the interclass separability in the penalty graph. It can be formulated using the Graph Embedding framework as explained in Section II. Thus, it can be extended using our framework to incorporate the data uncertainty using Eq. (18)-(20).

Figure 3 illustrates the performance of the original MFA and its uncertainty extension, i.e., GEU-MFA, for different combinations of reduced dimension d and k used in k-Nearest Neighbors (k-NN). We note that for small values of k and d, GEU-MFA performs better than the original method. For the extreme case ($k = 1, d = 1$), MFA has 52.8% accuracy compared to 77.1% for GEU-MFA. For higher values of (d, k), the performance of both approaches increase and they tend to perform similarly.

TABLE 1. Classification accuracy of MFA [4], RMFA [4], GEU-MFA-U, and GEU-MFA-S in the different datasets.

	noise	MFA	RMFA	GEU-MFA-U	GEU-MFA-S
Cancer	0%	0.858	0.851	0.866	**0.894**
	10%	0.833	0.870	0.884	**0.890**
	20%	0.806	0.825	0.835	**0.849**
Cifar2	0%	0.505	0.511	0.512	**0.520**
	10%	0.500	0.507	0.511	**0.513**
	20%	0.504	0.503	**0.506**	**0.506**
Yale B face	0%	0.910	0.913	**0.922**	0.921
	10%	0.901	0.902	0.905	**0.910**
	20%	0.892	0.896	0.901	**0.902**

In Figure 4, we show the performance of MFA, Regularized MFA (RMFA) and our variants of MFA as a function of the number of training samples on Cifar2. We note that incorporating uncertainty consistently yields a performance boost for both variants of uncertainty techniques compared to the original MFA. For smaller training data sizes, the supervised variant usually leads to slightly better results (less than 1%) than the unsupervised variant. When a higher number of training data is available, the unsupervised technique usually achieves the best accuracy.

In Table 1, we show the robustness of the standard MFA [4], Regularized MFA (RMFA) [4], and our proposed approach with both variants of uncertainty estimation, i.e., our MFA variant with unsupervised uncertainty variant (GEU-MFA-U) and our MFA variant with supervised uncertainty variant (GEU-MFA-S), on the three datasets with different additional noise levels. We repeat each experiment ten times and report the average accuracy achieved by each method. We note that the proposed methods outperform the original MFA for all noise levels. We also note that the accuracies of all the methods drop clearly when the noise level is higher. The supervised technique for estimating the uncertainty achieves the top performance except for Yale B Face dataset with no additional noise, where the best performance is achieved by GEU-MFA-U.

B. LDA

In Figure 5, we evaluate the performance of LDA, Regularized LDA (RLDA), our LDA variant with unsupervised uncertainty variant (GEU-LDA-U) and our LDA variant with supervised uncertainty variant GEU-LDA-S as a function of the number of training samples on Cifar2. We repeat each experiment ten times and report the mean and the variance of accuracies for all the training sizes. Similar to MFA, incorporating uncertainty yields a performance boost for both variants of uncertainty techniques compared to the original LDA. We also note that for higher number of training samples, the performance gap decreases. Both variants of uncertainty estimations achieve a similar performance for different training sizes.

TABLE 2. Classification accuracy of LDA [30], RLDA [4], RSLDA [23], ULDA [9], GEU-LDA-U, and GEU-LDA-S in the different datasets.

	noise	LDA	RLDA	RSLDA	ULDA	GEU-LDA-U	GEU-LDA-S
Cifar2	0%	0.523	0.541	0.511	0.505	**0.544**	0.535
	10%	0.497	0.538	0.516	0.501	0.542	**0.547**
	20%	0.523	0.545	0.510	0.498	0.541	**0.546**
Cancer	0%	0.932	**0.958**	0.882	0.528	0.951	0.950
	10%	0.896	**0.919**	0.858	0.541	0.917	0.918
	20%	0.895	**0.909**	0.829	0.505	0.904	0.901
Yale B	0%	0.856	0.869	0.851	0.871	**0.872**	0.871
	10%	0.849	**0.864**	0.827	0.859	0.863	0.862
	20%	0.838	0.853	0.839	0.852	**0.856**	0.855

We report the performance of LDA [30], regularized LDA [4], Robust Sparse Linear Discriminant Analysis (RSLDA) [23], Uncertain Linear Discriminant Analysis (ULDA) [9], GEU-LDA-U, and GEU-LDA-S on the three datasets for different noise levels in Table 2. We repeat each experiment ten times and report the average accuracy achieved by each approach. For the clean Cifar2 dataset, the best accuracy is achieved by GEU-LDA-U, while for the noisy Cifar2, GEU-LDA-S achieves the best results. The regularized LDA yields the best accuracy for Cancer and Yale B (noise=10%) datasets. One plausible explanation of this is that Cancer dataset is linearly separable dataset and using uncertainty might have made the problem harder. However, for the other two variants of Yale B dataset, the highest accuracy is achieved by GEU-LDA-U. Compared to the original LDA, the LDA variants obtained via the proposed framework are more robust to the presence of noise and yield higher accuracies.

V. CONCLUSION

In this work, we introduced a novel spectral-based dimensionality reduction framework called Graph Embedding with Data Uncertainty (GEU) that reformulates the Graph Embedding to consider input data uncertainties and artifacts. We model the uncertainty around each data point by a multivariate Gaussian distribution centered around the original sample and a covariance matrix characterizing the uncertainty of the corresponding sample along each feature dimension. Two techniques to generate the distribution of each data point were proposed based on the pair-wise distances between samples. Uncertainty introduces a regularization term that expands the rank of the scatter matrices and increases the number of available projection directions compared to the original subspace learning methods. We studied as special cases of the proposed framework the traditional subspace learning techniques LDA and MFA. The proposed framework was extensively evaluated over three datasets and it led to performance improvement compared to the original methods as well competing methods that consider uncertainty.

REFERENCES

[1] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics Intell. Lab. Syst.*, vol. 2, nos. 1–3, pp. 37–52, 1987.
[2] M. S. Park and J. Y. Choi, "Theoretical analysis on feature extraction capability of class-augmented PCA," *Pattern Recognit.*, vol. 42, no. 11, pp. 2353–2362, Nov. 2009.
[3] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1491–1497, Sep. 2013.
[4] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, Q. Yang, and S. Lin, "Graph embedding and extensions: A general framework for dimensionality reduction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 1, pp. 40–51, Jan. 2007.
[5] J. Bi and T. Zhang, "Support vector classification with input data uncertainty," in *Proc. Adv. Neural Inf. Process. Syst.*, 2004, pp. 161–168.
[6] C. Tzelepis, V. Mezaris, and I. Patras, "Linear maximum margin classifier for learning from uncertain data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2948–2962, Dec. 2018.
[7] K. Gajamannage, R. Paffenroth, and E. M. Bollt, "A nonlinear dimensionality reduction framework using smooth geodesics," *Pattern Recognit.*, vol. 87, pp. 226–236, Mar. 2019.
[8] Y. Pan, S. S. Ge, and A. Al Mamun, "Weighted locally linear embedding for dimension reduction," *Pattern Recognit.*, vol. 42, no. 5, pp. 798–811, May 2009.
[9] R. Saeidi, R. F. Astudillo, and D. Kolossa, "Uncertain LDA: Including observation uncertainties in discriminative transforms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 7, pp. 1479–1488, Jul. 2016.
[10] W. Zheng, C. Lu, Z. Lin, T. Zhang, Z. Cui, and W. Yang, "ℓ_1-norm heteroscedastic discriminant analysis under mixture of Gaussian distributions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 2898–2915, Oct. 2019.
[11] N. Vaswani, T. Bouwmans, S. Javed, and P. Narayanamurthy, "Robust subspace learning: Robust PCA, robust subspace tracking, and robust subspace recovery," *IEEE Signal Process. Mag.*, vol. 35, no. 4, pp. 32–55, Jul. 2018.
[12] R. Huang, Q. Liu, H. Lu, and S. Ma, "Solving the small sample size problem of LDA," in *Proc. Int. Conf. Pattern Recognit.*, vol. 3, 2002, pp. 29–32.
[13] V. Mygdalis, A. Iosifidis, A. Tefas, and I. Pitas, "Graph embedded one-class classifiers for media data classification," *Pattern Recognit.*, vol. 60, pp. 585–595, Dec. 2016.
[14] A. Iosifidis and M. Gabbouj, "Multi-class support vector machine classifiers using intrinsic and penalty graphs," *Pattern Recognit.*, vol. 55, pp. 231–246, Jul. 2016.
[15] D. Bouzas, N. Arvanitopoulos, and A. Tefas, "Graph embedded nonparametric mutual information for supervised dimensionality reduction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 5, pp. 951–963, May 2015.
[16] N. Passalis and A. Tefas, "Dimensionality reduction using similarity-induced embeddings," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3429–3441, Aug. 2018.
[17] L. Yang, S. Song, Y. Gong, H. Gao, and C. Wu, "Nonparametric dimension reduction via maximizing pairwise separation probability," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3205–3210, Oct. 2019.
[18] L. Wang and R.-C. Li, "Learning low-dimensional latent graph structures: A density estimation approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1098–1112, Apr. 2020.
[19] C. Örnek and E. Vural, "Nonlinear supervised dimensionality reduction via smooth regular embeddings," *Pattern Recognit.*, vol. 87, pp. 55–66, Mar. 2019.
[20] Ç. Aytekin, A. Iosifidis, S. Kiranyaz, and M. Gabbouj, "Learning graph affinities for spectral graph-based salient object detection," *Pattern Recognit.*, vol. 64, pp. 159–167, Apr. 2017.
[21] Y. Li, J. Chen, and L. Feng, "Dealing with uncertainty: A survey of theories and practices," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2463–2482, Nov. 2013.
[22] P. Lourenço, B. J. Guerreiro, P. Batista, P. Oliveira, and C. Silvestre, "Uncertainty characterization of the orthogonal Procrustes problem with arbitrary covariance matrices," *Pattern Recognit.*, vol. 61, pp. 210–220, Jan. 2017.
[23] J. Wen, X. Fang, J. Cui, L. Fei, K. Yan, Y. Chen, and Y. Xu, "Robust sparse linear discriminant analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 2, pp. 390–403, Feb. 2019.
[24] C.-N. Li, Y.-H. Shao, W. Yin, and M.-Z. Liu, "Robust and sparse linear discriminant analysis via an alternating direction method of multipliers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 915–926, Mar. 2020.
[25] Z. Yue, H. Yong, D. Meng, Q. Zhao, Y. Leung, and L. Zhang, "Robust multiview subspace learning with nonindependently and nonidentically distributed complex noise," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1070–1083, Apr. 2020.

[26] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 23, 2010, pp. 2496–2504.
[27] W. Zhao, R. Chellappa, and P. J. Phillips, "Subspace linear discriminant analysis for face recognition," Center Automat. Res., Univ. Maryland, College Park, MD, USA, Tech. Rep. CAR-TR-914, 1999.
[28] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep., 2009.
[29] K.-C. Lee, J. Ho, and D. J. Kriegman, "Acquiring linear subspaces for face recognition under variable lighting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 684–698, May 2005.
[30] H. Yu and J. Yang, "A direct LDA algorithm for high-dimensional data—With application to face recognition," *Pattern Recognit.*, vol. 34, no. 10, pp. 2067–2070, Oct. 2001.

FIRAS LAAKOM received the engineering degree from the Tunisia Polytechnic School (TPS), in 2018. He is currently pursuing the Ph.D. degree with Tampere University, Finland. He has coauthored three international journal articles and seven papers in international conferences and workshops. His research interests include machine learning, computer vision, learning theory, and computational intelligence.

JENNI RAITOHARJU (Member, IEEE) received the Ph.D. degree from the Tampere University of Technology, Finland, in 2017. She currently works as a Senior Research Scientist at the Finnish Environment Institute, Jyväskylä. She has coauthored 26 international journal articles and 42 papers in international conferences. She leads two research projects funded by the Academy of Finland focusing on automatic taxa identification. Her research interests include machine learning and pattern recognition methods along with applications in biomonitoring and autonomous systems. She is the Chair of the Young Academy Finland for the period 2019–2022.

NIKOLAOS PASSALIS received the B.Sc. degree in informatics, the M.Sc. degree in information systems, and the Ph.D. degree in informatics from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2013, 2015, and 2018, respectively. Since 2019, he has been a Postdoctoral Researcher with the Aristotle University of Thessaloniki, while from 2018 to 2019, he has also conducted postdoctoral research at the Faculty of Information Sciences, Tampere University, Finland. He has (co)authored more than 45 journal articles and 60 conference papers. His research interests include deep learning, information retrieval, time-series analysis, and computational intelligence.

ALEXANDROS IOSIFIDIS (Senior Member, IEEE) is currently a Professor at Aarhus University, Denmark. He has contributed to more than 30 research and development projects financed by EU, Finnish, and Danish funding agencies and companies. He has coauthored 92 articles in international journals and 120 papers in international conferences and workshops proposing novel machine learning techniques and their application in a variety of problems. He is a member of the IEEE Technical Committee on Machine Learning for Signal Processing. He is the Associate Editor-in-Chief of *Neurocomputing* journal covering the research area of neural networks and an Associate Editor of IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.

MONCEF GABBOUJ (Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from Purdue University, in 1986 and 1989, respectively. He was an Academy of Finland Professor, from 2011 to 2015. He is currently a Professor of information technology at the Department of Computing Sciences, Tampere University, Tampere, Finland. His research interests include big data analytics, multimedia content-based analysis, indexing and retrieval, artificial intelligence, machine learning, pattern recognition, nonlinear signal and image processing and analysis, voice conversion, and video processing and coding. He is a member of the Academia Europaea and the Finnish Academy of Science and Letters. He has served as the General Chair for IEEE SPS Conference and CAS Flagship Conference, ICIP and ISCAS, and ICME. He has served as an associate editor and a guest editor for many IEEE and other international journals.